

CLAIMS

1 1. An apparatus comprising:
2 at least one processor;
3 a memory coupled to the at least one processor;
4 a database table residing in the memory; and
5 a cardinality estimator residing in the memory and executed by the at least one
6 processor, the cardinality estimator estimating cardinality of an intermediate dataset that
7 satisfies a query to the database table in a manner that accounts for data skew in the
8 database table.

1 2. The apparatus of claim 1 further comprising a frequent values list residing in the
2 memory that contains a list of values in the database table, each value having a
3 corresponding frequency, wherein the cardinality estimator estimates the cardinality of
4 the intermediate dataset by determining whether a frequency corresponding to a value
5 exceeds a predetermined threshold, and if the frequency exceeds the predetermined
6 threshold, accounting for the corresponding value, and if the frequency does not exceed
7 the predetermined threshold, using a formula to estimate the cardinality of the
8 intermediate dataset, the formula accounting for data skew in the database table by
9 subtracting the frequency of all values above the predetermined threshold in the frequent
10 value table that satisfy the query from the total number of columns in the database table.

1 3. The apparatus of claim 2 wherein the cardinality estimator estimates the
2 cardinality Ca' of the intermediate dataset using the formula:

$$Ca' = P + M(1 - (1 - \frac{1}{M})^Y)$$

4 where

5 $M = Ca - (P+Q)$

6 P = number of distinct values in the frequent values list above the
7 predetermined threshold that satisfy the query;

Q = number of distinct values in the frequent values list above the
predetermined threshold that do not satisfy the query;

10 **Ca** = cardinality of the database table;

11 $Y = X - Fi;$

12 X = number of rows in the intermediate dataset; and

13 F_i = sum of frequencies of values in the frequent values list above the
14 predetermined threshold that satisfy the query.

1 5. A method for estimating cardinality of an intermediate dataset that results from
2 processing a database query on a database table, the method comprising the steps of:
3 (A) evaluating the query; and
4 (B) estimating cardinality of the intermediate dataset using a formula that
5 accounts for data skew in the database table.

1 6. The method of claim 5 wherein step (B) includes the steps of:
2 selecting a value in a frequent values list that contains a list of values in the
3 database table, each value having a corresponding frequency;
4 if the selected value has a corresponding frequency that exceeds a predetermined
5 threshold, incrementing the cardinality estimate by one; and
6 if the frequency does not exceed the predetermined threshold, using a formula to
7 estimate the cardinality of the intermediate dataset, the formula accounting for data skew
8 in the database table by subtracting the frequency of all values above the predetermined
9 threshold in the frequent value table that satisfy the query from the total number of
10 columns in the database table.

1 7. The method of claim 6 wherein the cardinality estimator estimates the cardinality
2 Ca' of the intermediate dataset in step (B) using the formula:

3
$$Ca' = P + M(1 - (1 - \frac{1}{M})^Y)$$

4 where

5
$$M = Ca - (P+Q)$$

6 P = number of distinct values in the frequent values list above the
7 predetermined threshold that satisfy the query;

8 Q = number of distinct values in the frequent values list above the
9 predetermined threshold that do not satisfy the query;

10 Ca = cardinality of the database table;

11
$$Y = X - Fi;$$

12 X = number of rows in the intermediate dataset; and

13 Fi = sum of frequencies of values in the frequent values list above the
14 predetermined threshold that satisfy the query.

- 1 8. A method for estimating cardinality of an intermediate dataset that results from
2 processing a database query on a database table, the method comprising the steps of:
3 (A) evaluating the query; and
4 (B) estimating the cardinality Ca' of the intermediate dataset using the formula:

5
$$Ca' = P + M(1 - (1 - \frac{1}{M})^Y)$$

6 where

7
$$M = Ca - (P+Q)$$

8 P = number of distinct values in the frequent values list above the
9 predetermined threshold that satisfy the query;

10 Q = number of distinct values in the frequent values list above the
11 predetermined threshold that do not satisfy the query;

12 Ca = cardinality of the database table;

13 $Y = X - Fi$;

14 X = number of rows in the intermediate dataset; and

15 Fi = sum of frequencies of values in the frequent values list above the
16 predetermined threshold that satisfy the query.

- 1 9. A program product comprising:
- 2 (A) cardinality estimator estimating cardinality of an intermediate dataset that
- 3 satisfies a query to a database table in a manner that accounts for data skew in the
- 4 database table; and
- 5 (B) computer-readable signal bearing media bearing the cardinality estimator.
- 1 10. The program product of claim 9 wherein the computer-readable signal bearing
- 2 media comprises recordable media.
- 1 11. The program product of claim 9 wherein the computer-readable signal bearing
- 2 media comprises transmission media.
- 1 12. The program product of claim 9 wherein the cardinality estimator evaluates a
- 2 frequent values list that contains a list of values in the database table, each value having a
- 3 corresponding frequency, wherein the cardinality estimator estimates the cardinality of
- 4 the intermediate dataset by determining whether a frequency corresponding to a value
- 5 exceeds a predetermined threshold, and if the frequency exceeds the predetermined
- 6 threshold, accounting for the corresponding value, and if the frequency does not exceed
- 7 the predetermined threshold, using a formula to estimate the cardinality of the
- 8 intermediate dataset, the formula accounting for data skew in the database table by
- 9 subtracting the frequency of all values above the predetermined threshold in the frequent
- 10 value table that satisfy the query from the total number of columns in the database table.

1 13. The program product of claim 12 wherein the cardinality estimator estimates the
2 cardinality Ca' of the intermediate dataset using the formula:

3
$$Ca' = P + M(1 - (1 - \frac{1}{M})^r)$$

4 where

5
$$M = Ca - (P+Q)$$

6 P = number of distinct values in the frequent values list above the
7 predetermined threshold that satisfy the query;

8 Q = number of distinct values in the frequent values list above the
9 predetermined threshold that do not satisfy the query;

10 Ca = cardinality of the database table;

11 $Y = X - Fi$;

12 X = number of rows in the intermediate dataset; and

13 Fi = sum of frequencies of values in the frequent values list above the
14 predetermined threshold that satisfy the query.

1 14. A program product comprising:
2 (A) a cardinality estimator that estimates cardinality of the intermediate using the
3 following formula:

4
$$Ca' = P + M(1 - (1 - \frac{1}{M})^Y)$$

5 where

6
$$M = Ca - (P+Q)$$

7 P = number of distinct values in the frequent values list above the
8 predetermined threshold that satisfy the query;

9 Q = number of distinct values in the frequent values list above the
10 predetermined threshold that do not satisfy the query;

11 Ca = cardinality of the database table;

12 $Y = X - Fi$;

13 X = number of rows in the intermediate dataset; and

14 Fi = sum of frequencies of values in the frequent values list above the
15 predetermined threshold that satisfy the query; and

16 (B) computer-readable signal bearing media bearing the cardinality estimator.

1 15. The program product of claim 14 wherein the computer-readable signal bearing
2 media comprises recordable media.

1 16. The program product of claim 14 wherein the computer-readable signal bearing
2 media comprises transmission media.

* * * * *